Probabilistic Latent Maximal Marginal Relevance

Shengbo Guo
ANU & NICTA
Canberra, Australia
shengbo.guo@nicta.com.au

Scott Sanner
NICTA & ANU
Canberra, Australia
scott.sanner@nicta.com.au

ABSTRACT

Diversity has been heavily motivated in the information retrieval literature as an objective criterion for result sets in search and recommender systems. Perhaps one of the most well-known and most used algorithms for result set diversication is that of Maximal Marginal Relevance (MMR). In this paper, we show that while MMR is somewhat ad-hoc and motivated from a purely pragmatic perspective, we can derive a more principled variant via probabilistic inference in a latent variable graphical model. This novel derivation presents a formal probabilistic latent view of MMR (PLMMR) that (a) removes the need to manually balance relevance and diversity parameters, (b) shows that specific definitions of relevance and diversity metrics appropriate to MMR *emerge* naturally, and (c) formally derives variants of latent semantic indexing (LSI) similarity metrics for use in PLMMR. Empirically, PLMMR outperforms MMR with standard term frequency based similarity and diversity metrics since PLMMR maximizes latent diversity in the results.

Categories and Subject Descriptors

H.3.3 [**Information Search and Retrieval**]: Retrieval Models

General Terms

Algorithms

Keywords

diversity, graphical models, maximal marginal relevance

1. INTRODUCTION

Maximal marginal relevance (MMR) [2] is perhaps one of the most popular methods for balancing relevance and diversity in set-based information retrieval and has been cited over 530 times[1] since its publication in 1998.

The basic idea of MMR is straightforward: suppose we have a set of items D and we want to recommend a small subset $S_k \subset D$ (where $|S_k| = k$ and $k \ll |D|$) relevant to a given query $\mathbf{q}$. MMR proposes to build S_k in a greedy manner by selecting s_j^* given $S_{j-1} = \{s_1^*, \ldots, s_{j-1}^*\}$ (where $S_j = S_{j-1} \cup \{s_j^*\}$) according to the following criteria

$$s_j^* = \operatorname*{arg\,max}_{s_j \in D \setminus S_{j-1}} [\lambda(\mathrm{Sim}_1(s_j, \mathbf{q})) - (1-\lambda) \max_{s_i \in S_{j-1}} \mathrm{Sim}_2(s_j, s_i)] \tag{1}$$

where $\mathrm{Sim}_1(\cdot,\cdot)$ measures the relevance between an item and a query, $\mathrm{Sim}_2(\cdot,\cdot)$ measures the similarity between two items, and the manually tuned $\lambda \in [0,1]$ trades off relevance and similarity. In the case of s_1^*, the second term disappears.

While MMR is a popular algorithm, it was specified in a rather ad-hoc manner and good performance typically relies on careful tuning of the λ parameter. Furthermore, MMR is agnostic to the specific similarity metrics used, which indeed allows for flexibility, but makes no indication as to the choice of similarity metrics for Sim_1 and Sim_2 that are compatible with each other and also appropriate for good performance.

In the next section, we address these concerns by taking a more principled approach to set-based information retrieval via maximum *a posteriori* probabilistic inference in a latent variable graphical model of marginal relevance (PLMMR). As an elegant and novel contribution, we note that natural relevance and diversity metrics *emerge* from this derivation (with no analogous manually tuned λ parameter) and that these metrics *also* formally motivate variants of similarity metrics used in latent semantic indexing (LSI) [3].

2. PROBABILISTIC LATENT MMR

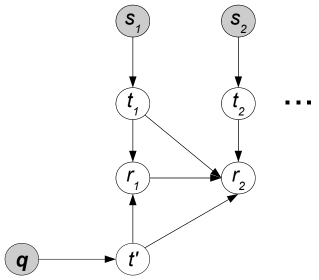

Figure 1: Graphical model used in PLMMR.

We begin our discussion of PLMMR by introducing a graphical model of (marginal) relevance in Figure 1. Shaded nodes represent observed variables while unshaded nodes are

[1] According to Google Scholar.

SIGIR'10, July 19–23, 2010, Geneva, Switzerland.
Copyright 2010 ACM 978-1-60558-896-4/10/07 ...$10.00.

latent; we do not distinguish between variables and their assignments. The observed variables are the vector of query terms $\mathbf{q}$ and the selected items $s_1 \in D$ and $s_2 \in D$. For the latent variables, let T be a discrete topic set; variables $t_1 \in T$ and $t_2 \in T$ respectively represent topics for s_1 and s_2 and $t' \in T$ represents a topic for query $\mathbf{q}$. $r_1 \in \{0, 1\}$ and $r_2 \in \{0, 1\}$ are variables that indicate whether the respective selected items s_1 and s_2 are relevant (1) or not (0).

The conditional probability tables (CPTs) in this discrete directed graphical model are defined as follows. $P(t_1|s_1)$ and $P(t_2|s_2)$ represent topic models of the items and $P(t'|\mathbf{q})$ represents a topic model of the query. There are a variety of ways to learn these topic CPTs based on the nature of the items and query; for an item set D consisting of text documents and a query that can be treated as a text document, a natural probabilistic model for $P(t_i|s_i)$ and $P(t'|q)$ can be derived from Latent Dirichlet Allocation (LDA) [1]. Finally, the CPTs for relevance r_i have a very natural definition:

$$P(r_1|t', t_1) = \begin{cases} 1 & \text{if } t_1 = t' \\ 0 & \text{if } t_1 \neq t' \end{cases}$$

$$P(r_2|t', r_1 = 0, t_1, t_2) = \begin{cases} 1 & \text{if } (t_2 \neq t_1) \wedge (t_2 = t') \\ 0 & \text{if } (t_2 = t_1) \vee (t_2 \neq t') \end{cases}$$

Simply, s_1 is relevant if its topic $t_1 = t$ (the query topic). s_2 is relevant with the same condition and the addition that if s_1 was irrelevant ($r_1 = 0$), then topic t_2 for s_2 should also not match t_1. Following the *click-chain model*, weassume the user *only* examines s_2 if s_1 was irrelevant ($r_1 = 0$).

Let us assume that like MMR we use a greedy item set selection algorithm and we have already selected $s_1 = s_1^*$. Now given $S_1 = \{s_1^*\}$, we want to select s_2 in order to maximize its *marginal relevance* w.r.t. $\mathbf{q}$ given S_1, formally defined as $MR(S_1, s_2, \mathbf{q})$ and derived as a query in the graphical model:

$$\begin{aligned} s_2^* &= \operatorname*{arg\,max}_{s_2 \in D \setminus S_1} MR(S_1, s_2, \mathbf{q}) = \operatorname*{arg\,max}_{s_2 \in D \setminus \{s_1^*\}} P(r_2|s_1^*, s_2, \mathbf{q}) \\ &= \operatorname*{arg\,max}_{s_2 \in D \setminus \{s_1^*\}} \sum_{t_1, t_2, t'} P(r_2|r_1 = 0, t_1, t_2, t')P(t_1|s_1^*) \\ &\qquad\qquad P(r_1 = 0|t_1, t')P(t_2|s_2)P(t'|\mathbf{q}) \\ &= \operatorname*{arg\,max}_{s_2 \in D \setminus \{s_1^*\}} \underbrace{\left(\sum_{t'} P(t'|\mathbf{q})P(t_2 = t'|s_2) \right)}_{\text{relevance}} - \\ &\qquad \underbrace{\left(\sum_{t'} P(t'|\mathbf{q})P(t_1 = t'|s_1^*)P(t_2 = t'|s_2) \right)}_{\text{diversity}} \qquad (2) \end{aligned}$$

The basic insight leading to this fascinating result is the exploitation of the indicator structure of the relevance variables r_1 and r_2 to make convenient variable substitutions.

We note that in this special case for $MR(S_1, s_2, \mathbf{q})$, a very natural mapping to the MMR algorithm in (1) when $\lambda = 0.5$ has *emerged* automatically from the derivation that maximized MR. This derivation automatically balances relevance and diversity without an analogous λ *and* it suggests very specific (and different) relevance and diversity metrics, both effectively variants of similarity metrics used in latent semantic indexing (LSI) [3]. To make this clear, we examine the *relevance* metric Sim_1^{PLMMR} given by PLMMR where we let $\mathbf{T}'$ and $\mathbf{T_2}$ be respective topic probability vectors for query $\mathbf{q}$ and item s_2 with vector elements $\mathbf{T}'_i = P(t' = i|\mathbf{q})$ and $\mathbf{T_2}_i = P(t_2 = i|s_2)$ and using $\langle \cdot, \cdot \rangle$ for the inner product:

$$\text{Sim}_1^{PLMMR}(\mathbf{q}, s_2) = \sum_{t'} P(t'|\mathbf{q})P(t_2 = t'|s_2) = \langle \mathbf{T}', \mathbf{T_2} \rangle.$$

A similar analysis gives *diversity* metric $\text{Sim}_2^{PLMMR}(s_1, s_2)$, yielding a variant LSI similarity metric *reweighted* by the query topic probability $P(t'|\mathbf{q})$. This points out the important correction to MMR that item set diversity should be query-relevant! Given these definitions of Sim_1^{PLMMR} and Sim_2^{PLMMR}, we can now substitute these into the MMR algorithm defined in (1) to arrive at a definition of PLMMR.

Table 1: Weighted subtopic loss (WSL) of three methods using all words and first 10 words. Standard error estimates are shown for PLMMR-LDA.

Method	WSL (first 10 words)	WSL (all words)
MMR-TF	0.555	0.534
MMR-TFIDF	0.549	0.493
PLMMR-LDA	**0.458 ± 0.0058**	**0.468 ± 0.0019**

3. EXPERIMENTAL COMPARISON

We report experiments on a subset of TREC 6-8 data focusing on diversity. We follow the same experimental setup as [5] who measure the weighted subtopic loss (WSL) of recommended item sets where in brief, WSL gives higher penalty for not covering popular subtopics. We do not compare directly to [5] as their method was supervised while MMR and PLMMR are inherently unsupervised.

Standard query and item similarity metrics used in MMR applied to text data include the cosine of the term frequency (TF) and TF inverse document frequency (TFIDF) vector space models [4]. We denote these variants of MMR as MMR-TF and MMR-TFIDF. PLMMR specifically suggests the use of LSI-based similarity metrics defined in the last section; thus, we use LDA to derive these models, referring to the resulting algorithm as PLMMR-LDA. LDA was trained with $\alpha = 2.0, \beta = 0.5, |T| = 15$; we note the results were not highly sensitive to these parameter choices.

Average WSL scores are shown in Table 1 on the 17 queries examined by [5]. We use both full documents and also just the first 10 words of each document. For both MMR algorithms, the best performing $\lambda = 0.5$ is shown. We note that due to the power of the latent topic model and derived similarity metrics, PLMMR-LDA is able to perform better than MMR with standard TF and TFIDF metrics and *without* a λ parameter to be tuned. In addition, PLMMR-LDA works very well with *short documents* since intrinsic document and query similarities are *automatically* derived from the latent PLMMR relevance and diversity metrics.

Acknowledgements

NICTA is funded by the Australian Government as represented by the Department of Broadband, Communications and the Digital Economy and the Australian Research Council through the ICT Centre of Excellence program.

4. REFERENCES

[1] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent Dirichlet allocation. *JMLR*, 3:993–1022, 2003.

[2] J. Carbonell and J. Goldstein. The use of MMR, diversity-based reranking for reordering documents and producing summaries. In *SIGIR*, 335–336. 1998.

[3] S. Deerwester, S. T. Dumaisand, G. W. Furnas, T. K. Landauer, and R. Harshman. Indexing by latent semantic analysis. *JASIS*, 41:391–407, 1990.

[4] G. Salton and M. McGill. *Introduction to modern information retrieval.* McGraw-Hill, 1983.

[5] Y. Yue and T. Joachims. Predicting diverse subsets using structural SVMs. In *ICML*, 1224–1231, 2008.